More simulation hypothesis hullabaloo, a refutation redux

Benjamin James

December 4, 2025

The simulation hypothesis has always depended on a certain asymmetry of confidence. Physics is still groping toward a final account of reality, while speculative philosophy confidently announces that reality is "probably" a computer program. As long as physics appears incomplete, the simulation hypothesis can parasitize that uncertainty. What has begun to change in the last few years is that the direction of pressure is reversing. The closer physics gets to a genuine theory of everything, the less room there is for the simulation hypothesis to occupy. My earlier paper argued that once you take adaptive emergence, thermodynamic constraints, and recursive irreducibility seriously, the hypothesis is either physically impossible or logically meaningless. The undecidability work by Faizal, Krauss, and collaborators reaches the same destination from a different starting point. If our ultimate physics necessarily includes non-algorithmic structure, then no Turing-equivalent device can ever host a complete copy of our universe. They frame this as a consequence of Gödel, Tarski, and Chaitin. I agree with their core result, but I think the implications are stronger than they themselves are willing to draw. Once you integrate their undecidability constraints with the ontological, thermodynamic, and epistemic problems I previously outlined, the simulation hypothesis is not merely an unbuildable technology. It is not a coherent description of any possible world.

Their paper begins from a natural assumption that whatever our final quantum gravity theory looks like, it will take the form of a formal system $\mathcal{F}_{QG} = (\mathcal{L}_{QG}, \Sigma_{QG}, \mathcal{R}_{alg})$. Here, $\mathcal{L}_{QG}$ is a first-order language rich enough to talk about fields, curvature, states, and amplitudes. Σ_{QG} is a finite or at least recursively enumerable set of axioms encoding the basic principles. $\mathcal{R}_{alg}$ is an effective set of inference rules. The key point is that $\mathcal{F}_{QG}$ is supposed to be algorithmic. There should be, in principle, a procedure that enumerates all its theorems. Spacetime, in this view, is not a primitive arena but an emergent construct inside the models of $\mathcal{F}_{QG}$. All the

usual conceptual machinery of “it from bit” and emergent geometry is placed on top of this formal substrate.

Once you grant this framing, Gödel and his successors come for free. Because $\mathcal{F}_{QG}$ is consistent, finitely axiomatized, and arithmetically expressive, its set of theorems, $\mathrm{Th}(\mathcal{F}_{QG})$, are a strict subset of the set of statements that are actually true in its intended model, $\mathrm{True}(\mathcal{F}_{QG})$. Gödel’s first incompleteness theorem guarantees the existence of true but unprovable sentences; his second theorem blocks any internal proof of the system’s own consistency. Tarski’s undefinability theorem blocks an internal truth predicate that captures “φ is true” for all φ. Chaitin’s information-theoretic incompleteness gives a bound, $K_{\mathcal{F}}$, such that no statement of Kolmogorov complexity greater than $K_{\mathcal{F}}$ can be settled within the system. Faizal and coauthors interpret these as more than curiosities of mathematical logic. They argue that in any algorithmic theory of quantum gravity, there will be physically meaningful questions about specific black hole microstates, thermalization of many-body systems, or renormalization flows that are in principle undecidable by the formal machinery of $\mathcal{F}_{QG}$, even though they have definite truth values in the world.

Their proposed resolution is to enlarge $\mathcal{F}_{QG}$ into a “meta-theory of everything”,

$$\mathcal{M}_{ToE} = \left(\mathcal{L}_{QG} \cup \{T\}, \Sigma_{QG} \cup \Sigma_T, \mathcal{R}_{alg} \cup \mathcal{R}_{nonalg}\right)$$

An external truth predicate $T(x)$ and a non-effective inference mechanism $\mathcal{R}_{nonalg}$ certify precisely those statements that Gödel, Tarski, and Chaitin forbid $\mathcal{F}_{QG}$ from handling on its own. The axiom set Σ_T is not recursively enumerable. The resulting Th_T, the set of T-true sentences, is not computable. In physical terms, they suggest that certain Planck-scale truths, certain questions about microstates, phase structure, and quantum collapse, are real features of the universe that cannot be captured by any algorithmic theory but can nonetheless be accessed by some non-algorithmic “understanding” encoded in $T(x)$. Whether one agrees with their Penrose-style account of human cognition is secondary. The structural point is clear; if our universe actually instantiates something like $\mathcal{M}_{ToE}$, then any complete theory of it must contain irreducibly non-computable content.

They then turn the same machinery on the simulation hypothesis. Any genuine simulation, they note, is assumed to be an algorithmic, finite program running on some computational

substrate. At best, such a program could emulate the algorithmic slice $\mathcal{F}_{QG}$ of our physics, but it cannot incorporate the non-recursive truth predicate $T(x)$ or the non-effective rules $\mathcal{R}_{nonalg}$ that define $\mathcal{M}_{ToE}$. Any simulation would systematically omit precisely those non-computable structures that their analysis shows to be necessary for a complete and consistent account of quantum gravity. Therefore, they conclude, a simulated universe can never be physically equivalent to ours. If reality contains non-algorithmic structure at the deepest level, and simulations are by definition algorithmic, then the strong simulation hypothesis, the claim that our universe is literally the output of a finite program on a Turing-equivalent machine, is ruled out in principle.

That conclusion is damaging. It cleanly removes the most popular, digital-physics variant of the hypothesis. However, it leaves two escape hatches that simulation proponents can exploit if they are determined to keep the idea alive. The first is to abandon strict digitalism and claim that the "simulating entity" is not algorithmic at all, but some higher-order non-computational system that happens to instantiate $\mathcal{M}_{ToE}$ directly. The second is to retreat into metaphoric language and insist that "simulation" need not mean algorithmic emulation, but any case where one universe arises from another in dependence on its structure. Faizal and coauthors gesture at the first route when they talk about non-algorithmic understanding, but they do not fully close it. They simply say that reality cannot be a computer simulation. The Bostrom-style discourse they are pushing against, however, is quite capable of mutating in response. It can replace "computer" with "meta-physical process" and continue. To finish the job, one has to show that the simulation hypothesis fails not only as an engineering project, but as a coherent ontological category.

My earlier critique approached the problem from four angles that remain untouched by their undecidability analysis. First, there is the problem of the simulating entity itself. If you posit a simulator S that exists in some higher-order reality and generates our universe U, you immediately face an ontological disparity. Either S obeys the same fundamental principles as U, or it does not. If $P(S) = P(U)$, then S is not metaphysically "above" U at all; it is simply another physical system in the same ontological stratum. Calling U a "simulation" of S's world then adds nothing over calling it another universe with the same laws. If $P(S) \neq P(U)$,

S does not share the same dynamical structure, thermodynamic constraints, or undecidability profile. In that case, it has no straightforward way to generate an exact copy of U's emergent behavior. It could, at best, approximate some aspects of U in a lower-fidelity model. Either way, the clean hierarchical picture of "base reality" and "simulated reality" collapses.

Second, there is the problem of adaptive, recursive irreducibility. The observable universe is not a static dataset but an adaptive process with continual negotiation of possibilities via feedback, entropy management, and choice. The future of such a system is not compressible into a shortcut function of its initial conditions. Its evolution is computationally irreducible. Any would-be simulator must either precompute all possible trajectories of an adaptive system, which requires storing an information content equal to the full entropy of the universe, or it must track the system's evolution step by step at full scale. In the first case, the "simulation" becomes identical in complexity to the universe it emulates. In the second, it ceases to be a simulation in any meaningful sense and becomes just another instantiation of the same process. The unobtrusive program running a lightweight emulation in the background, immune to the universe's constraints, is nowhere to be found.

Third, there are brute thermodynamic constraints. Any physical system that performs computation must pay Landauer's cost. Each logically irreversible operation generates at least $k_B T \ln 2$ of heat at temperature T, where k_B is Boltzmann's constant. To track every degree of freedom in a universe-sized system at even modest resolution requires more bit operations per unit time than any finite physical substrate could support without collapse. A full-fidelity simulation of our universe's thermodynamic behavior must itself be a thermodynamic system of comparable scale. The energy and entropy flows cannot simply disappear "off-stage." Faizal and coauthors acknowledge the importance of thermodynamics in computation elsewhere, but in their anti-simulation argument they lean most heavily on undecidability rather than on energy accounting. The result is that they close the door on idealized Turing simulations but leave open a gap for metaphysical simulators with unspecified resources.

Finally, there is the epistemic closure problem. A hypothesis is scientifically meaningful only if it makes contact with possible experience, if there exists, in principle, some observation that could confirm or disconfirm it, but any simulation that preserves the physics we actually observe inherits a dilemma. If there exists some measurable signature that distinguishes "simulated" from "base" reality, then the simulation hypothesis must commit to concrete, testable deviations from our current theories. That puts us back in the realm of normal physics, not metaphysical speculation. If no such signature is possible, then "we are in a simulation" does not change a single prediction about any observation or experiment. It is empirically inert. The hypothesis then reduces to a rephrasing of "reality exists" with unnecessary extra entities attached. Combining this epistemic vacuity with the logical and physical difficulties above shows that the simulation hypothesis is not merely false; it fails to even meet the minimal criteria of a coherent scientific or philosophical claim.

What the undecidability work adds, and what makes it a useful complement to my earlier objections, is a sharpened statement about the limits of any algorithmic representation of physical law. If you accept their $\mathcal{F}_{QG}$ / $\mathcal{M}_{ToE}$ distinction, then you have to accept that no Turing machine, no matter how powerful, can ever host a complete universe. At best it can emulate a computable fragment and systematically miss any non-algorithmic meta-structure that is required for consistency. This plugs the most obvious hole in my earlier argument. The claim that no finite computational substrate can host a faithful copy of an adaptive, emergent universe is no longer just an appeal to complexity and thermodynamics. It is a statement about the mismatch between recursive theories and the truths they aim to describe.

The decisive step, however, comes when you integrate their result into the ontological disparity analysis instead of letting it stand alone. Suppose, conceding their point, that our universe requires a non-algorithmic meta-layer $\mathcal{M}_{ToE}$. The simulation hypothesis now has only one move, to assert that the simulating entity S also instantiates $\mathcal{M}_{ToE}$, and thus can host a universe with the same blend of computable and non-computable structure. However, once you say that, you have erased the metaphysical distance between S and U. There is no longer an "algorithmic device" over here and a "rich physical universe" over there.

There are simply two systems each instantiating the same non-algorithmic theory of everything. The relation between them cannot be characterized as simulation in any robust sense. It is, at best, a relation of one universe giving rise to another. At that point it is cleaner to drop the simulation vocabulary altogether and admit that we are simply talking about cosmology.

A necessary and sufficient refutation flows naturally out of this layered picture. First, any purely algorithmic simulation of a complete universe is impossible, by the combined weight of Gödel–Tarski–Chaitin and computational irreducibility. Second, any non-algorithmic "simulation" that fully instantiates our physics ceases to be a simulation and collapses into an ontologically peer universe. Third, any version of the simulation hypothesis that is empirically distinguishable from the claim "our universe is real" is just a speculative physical model in disguise and therefore falls under ordinary scientific scrutiny. Fourth, any version that is not empirically distinguishable adds no explanatory or predictive content and is therefore not a meaningful hypothesis. There is no formulation left that does what its proponents want, to place our reality in a subordinate, derivative, or "merely virtual" position relative to some higher-order computational substrate.

The more carefully we understand the logical limits of formal systems, the thermodynamic limits of computation, and the adaptive, irreducible character of physical reality, the less plausible the simulation frame becomes. What once seemed like a bold inversion, "we are not the top level, we are somebody else's process," turns out to be a confusion between representation and instantiation, between algorithm and emergence, and between metaphor and ontology. A universe that contains undecidable truths, non-algorithmic structure, and irreducible adaptive processes cannot be hosted by a finite computation. A system that fully instantiates such a universe cannot be cleanly distinguished from that universe in ontological rank. A hypothesis that cannot be tested, cannot guide expectation, and cannot survive contact with these constraints is not a live competitor in our attempt to understand reality.

If anything, the undecidability results from Faizal, Krauss, and others strengthens the coherence-first view I have been developing. Reality is not a subordinate layer in somebody's

information-processing stack. It is the adaptive, self-maintaining field in which any such stacks must themselves be embedded. Coherence is not an illusion produced by higher-level computation; computation is a special case of coherence dynamics. The simulation hypothesis tried to flip that order, making being contingent on some deeper bit-level. What recent work shows, and what a fully integrated refutation makes explicit, is that this inversion cannot be made precise. There is no possible world in which our universe is both as rich as it is and also merely the output of someone else's program. Once those conditions are laid side by side without equivocation, the hypothesis dies on its own terms.

References

Faizal, M., Krauss, L., Marino, F., & Shabir, A. (2025). Consequences of Undecidability in Physics on the Theory of Everything. Arxiv. https://arxiv.org/html/2507.22950v1

James, B. (2025). The Logical And Physical Impossibility Of The Simulation Hypothesis. PhilPapers. https://philpapers.org/rec/JAMTLA-4